Consumerism, Contradictions, Counterfactuals: Shaping the Evolution of Safety Engineering

John A McDermid, Zoë Porter, Yan Jia

Assuring Autonomy International Programme, University of York

Abstract *This paper takes the (perhaps unusual) view that consumerism has helped to drive improvements in safety over the years. However, the successes in terms of the availability of (safe) goods and services, e.g. cars and cheap air transport, present contradictions (or ironies) in terms of the subsequent impact on the environment which ultimately has a deleterious effect on safety and well-being. These contradictions suggest the need to re-frame safety engineering. The paper proposes an approach based on the notion of well-being and discusses how counterfactuals might play a role in analysing and communicating about safety concerns.*

1 Introduction

In several domains, sustained improvements to system safety have been achieved over many decades. This is perhaps particularly obvious in road vehicles and air transport, but the trend can be seen more generally. Some of the credit for this is due to good safety engineering and safety culture. But customer pressure is also a factor as safety has become essential to the sales of some products or services. This combines with other factors, such as progressive reductions in unit prices, to enable widespread availability of safe products and services. But it also comes at a cost. The most obvious is the environmental impact of increased road traffic and air travel which, in turn, has a negative impact on human safety. Because of ironies or contradictions such as these – whereby improvements to safety lead indirectly to greater physical risk – we propose a re-framing of safety engineering and assurance to look much more broadly and holistically at hazards and impacts. We also propose a re-framing of safety engineering and assurance to include (and codify) the intended benefit from the system as well as the possible harm. This widening of the safety and assurance landscape aligns with the precepts of 'Ethically Aligned Design' (IEEE 2019), and the principles of Responsible Research and Innovation (Owen et al. 2013; Von Schomberg 2013), but we seek here to define a more tangible approach to defining and assessing risk, more strictly a balance of risk and benefit.

Aligned with the re-framing of safety as a concern we also consider the role of counterfactual reasoning in the evaluation, communication, and mitigation of risk. The term ‘counterfactual’ has a long history in philosophy and related disciplines. A counterfactual statement can be defined loosely as being about “what was not, or is not, but could or would have been” and it is often expressed as subjunctive conditional: “if x had/had not occurred, then y would/would not have occurred” (Starr 2021). Although ‘counterfactual’ is now used in artificial intelligence (AI) to refer to a particular form of explanation method, our focus here is rather different. We consider several potential roles for counterfactual reasoning in this re-framing of safety engineering, e.g. supporting accident analysis.

The rest of the paper is structured as follows. Section 2 considers the relationship between *consumerism* and historical trends towards safer products and services. Section 3 considers the *contradictions* or ironies that arise from the successful reduction in cost, and improvements in safety, of products and services, particularly those that are sold on the mass market. Section 4 proposes a progressive re-framing of safety engineering which embraces the notion of well-being, codifies benefit as well as harm, and considers longer-term impacts such as environmental damage. The intent is that this resonates more fully with ethical and societal concerns and expectations about high-integrity systems, and that these ideas will help to shape the evolution of safety as a concern, and safety engineering as a discipline, to give a practical basis for ethically aligned design and systems engineering. Within this, we explore the potential role of counterfactual analysis in understanding and communicating about hazards and risks. In Section 5, we discuss how the proposed re-framing identifies and addresses issues that are not covered in current enlargements of safety engineering, such as Safety II (Hollnagel 2018), and new ethical standards for engineers, although the ideas presented here are complementary to those developments. We also raise open and unanswered questions, such as formalisations of ‘well-being’ and distribution of risk, emphasising the importance of multidisciplinary research in this evolving concern. Finally, we consider what steps might need to be taken to enable these broad concepts to influence real-world engineering.

2 Consumerism and the Achievement of Safety

Mature industries, such as aerospace and automotive, have seen sustained reductions in accident levels and fatalities over many years albeit with some geographical variation. In our view this is *in part* due to good safety engineering and safety management, but there are also other influences, such as the (implied) pressures on manufacturers to achieve societally acceptable levels of risk, the cost of recalls and potential reputational damage, which drive improvements to safety. Here we highlight one influence that seems to us to be particularly significant – consumerism. Consumerism has two definitions, or facets, which we might characterise as:

- The protection or promotion of the interests of customers and
- The preoccupation of society with the acquisition of goods or services.

Whilst these might seem somewhat contradictory, these facets *work together* when it comes to the impact on safety. We illustrate this by considering the very different effects of consumerism in two sectors: air transport and cars.

2.1 Air Transport and Tourism

One of the most famous graphics showing how air travel has become safer over the years is from Boeing's annual aviation statistics summary (Boeing 2020). Figure 1 shows the data from the late 1950s onwards. From a safety engineering perspective, this indicates (although it doesn't prove) the long-term effectiveness of safety analysis and management (notwithstanding the issues surrounding the Boeing 737 MAX). In particular, the approach to analysing accidents has meant that the industry has understood the underlying causes of accidents, e.g. "power structures" in cockpits, and introduced specific remedies such as approaches to crew/cockpit resource management (CRM) to reduce problems of communication and losses of situational awareness.

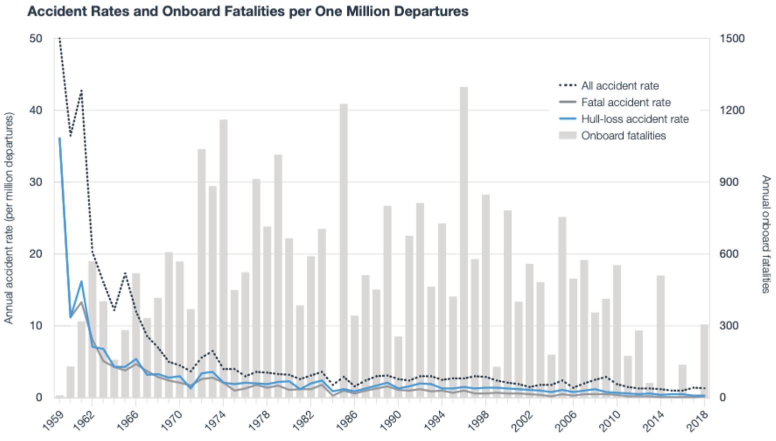


Fig. 1. Aircraft Accident Statistics (Boeing 2020)

Worldwide, aircraft departures have grown substantially over the years. They stood at over 35 million flights per annum in 2019, although this has dropped dramatically due to Covid-19 (World Bank 2021). Considering Figure 1, if the accident rate had remained at 1960s levels, then theoretically there would be roughly one aircraft accident per day. Of course, this is just hypothetical – given the impact of such accidents on public perception, and the impact of court cases and compensation claims, such an accident rate would very likely cause aircraft usage to drop dramatically, indeed to fundamentally transform the air transport sector.

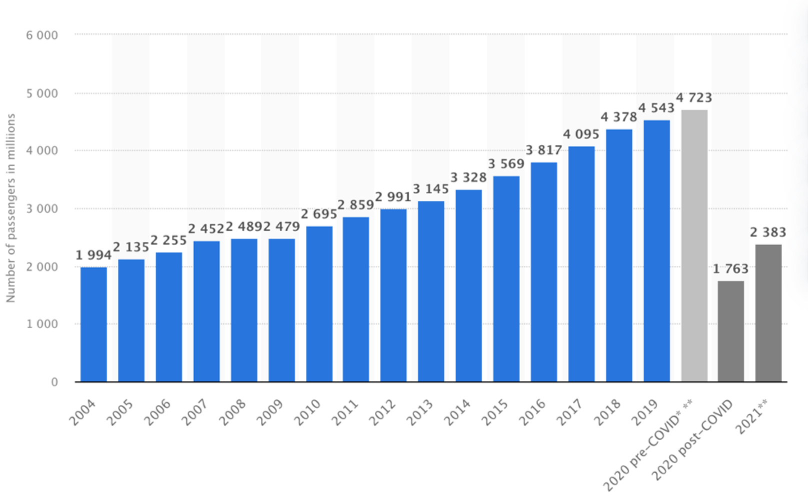


Fig. 2. Aircraft Passenger Numbers (Statista 2021)

So, what has driven this growth in traffic? The answer appears to be tourism which, of course, is an example of consumerism where people are interested in acquiring services – in this case cheap (international) holidays – resulting in the growth in passenger numbers illustrated in Figure 2. Air travel has changed from being a privilege to being commonplace, and an analysis suggests that tourism had become the world’s largest industry by 1984 (Lyth and Dierikx 1994) and it stood at over 10% of *global* Gross Domestic Product (GDP), pre-pandemic.

Whether or not aircraft safety would have improved so much without the pressures of consumerism is, of course, unknowable; we cannot show cause and effect. However, the counterfactual is clear – without the safety improvements, the traffic growth and scale of international tourism would not have occurred as society would not have embraced air travel if the apparent risks of flying were so high.

2.2 Cars and Shifting Expectations

The data from the automotive sector also show a significant downward trend in accidents over the years. Figure 3 is from a Department for Transport (DfT) summary for Great Britain (GB) over a 40-year period (Department for Transport 2020). As with air transport, this downward trend is against an increase in traffic volumes – with a fatal accident every 4.9 billion vehicle miles in 2019 against one every 7.1 billion miles in 2009. However, what is perhaps more telling is the improvements in vehicles that have occurred over many decades.

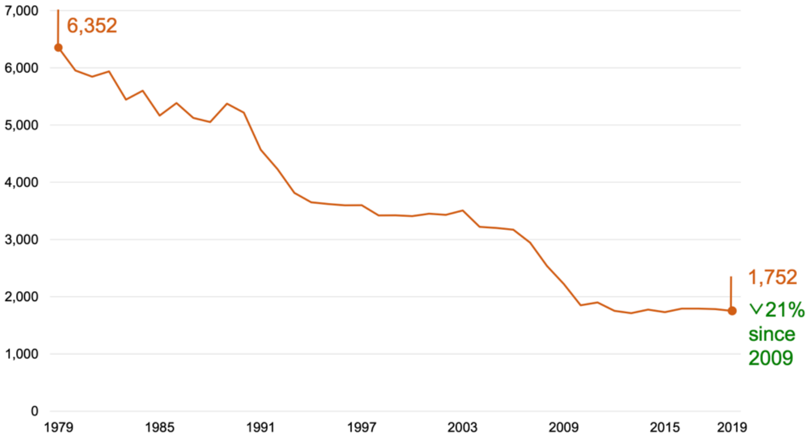


Fig. 3. Fatalities in Reported Road Accidents, GB 1979-2019 (DfT 2020)

First, it is worthwhile making some observations about consumerism. To be successful, products often need differentiators, or unique selling points (USPs), that set them apart from the competition. There are also minimum expectations on products and if these are not met by a particular product then it might not be successful, despite the presence of some attractive USPs. These minimum expectations (sometimes referred to as "table stakes" based on the use of the term in gambling) can change over time – indeed something that was once a USP can become a "table stakes" feature. This can be seen for cars, including for safety features, which is illustrated by the following partial timeline extracted from one produced by the UK Automobile Association (AA 2021)):

1911 – rear view mirrors.
1921 – headrests.
1951 – airbags.
1952 – crumple zones.
1963 – inertia reel seatbelts.

1978 – anti-lock braking.
1991 – side impact protection systems.
1995 – electronic stability control.
2000 – lane departure warning systems.
2008 – autonomous emergency braking.
2010 – pedestrian detection system.

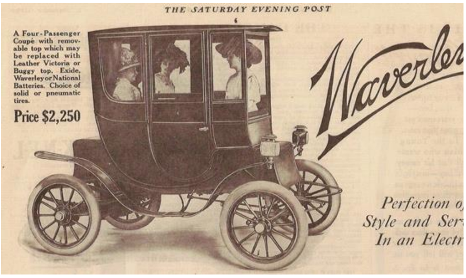
Many of these early innovations have become expected and some, e.g. seat belts, are now required by regulation. There is also a significant shift from “passive” safety, e.g. headrests, through to more “active” safety systems such as autonomous emergency braking. One can also add many items to this list that were once seen as “luxuries”, e.g. reversing cameras, which are now fitted on many vehicles.

Despite a similar effect on safety, the trends and impact of consumerism in the automotive sector is very different from than in air transport. Aircraft purchase (or lease) is the province of the professional and consumer pressure for safe, cheap services is indirect. Cars are (often) an individual purchase and the availability of safety features on some vehicles pushes the manufacturers to provide similar capabilities for fear of losing sales to other brands. Also, customer attitudes are important – it is not uncommon for people to say: “I won’t buy it unless it has X” (where X is some safety feature, e.g. anti-lock brakes). Finally, the (European) New Car Assessment Programme (Euro NCAP)[1] serves to keep safety in the public eye and a factor when purchasing a new vehicle. From a safety perspective this is good news as vehicle manufacturers strive to improve their NCAP rating, and there are now trends to protect vulnerable road users (VRUs).

There is another perspective on consumerism in the automotive sector which is relevant to our re-framing of safety. Electric cars first appeared as early as the 1830s, and were quite widespread by 1900, particularly in cities with good availability of electricity supplies (Department of Energy 2014). At the same time, cars based on the internal combustion engine were dirty and smelly – presaging today’s problems – and had other challenges, including being hard to drive. But the availability of cheap fuel and cheaper cars (a Model T Ford was about a third of the price of a similar electric vehicle) meant that more consumers could acquire goods by buying vehicles with internal combustion engines. Figures 4 and 5 are examples of advertisements that reflect the difference in price of electric vehicles and those with internal combustion engines; although the adverts aren’t strictly contemporaneous the difference in price ($2,250 vs $360) is indicative of the competitive problems of electric vehicles. Consequently, electric vehicles had almost disappeared by the 1930s. Again, a plausible counterfactual can be made – without the availability of cheap cars using internal combustion engines and cheap fuel, city transport would not have moved away from electric vehicles (see

[1] See: https://www.euroncap.com/en

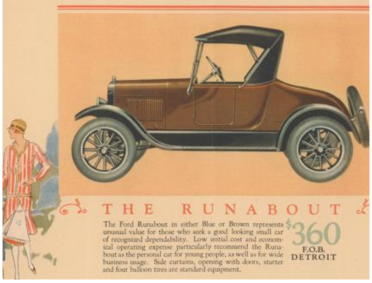
section 3.2 for a discussion of the wider implications of this change in propulsive power).



A Four-Passenger Coupé with removable top which may be replaced with Leather Victoria or Buggy top. Exide, Waverley or National Batteries. Choice of solid or pneumatic tires.

Price $2,250

Fig. 4. Electric Car advert circa 1900



The Ford Runabout in either Blue or Brown represents unusual value for those who seek a good looking small car of recognized dependability. Low initial cost and economical operating expense particularly recommend the Runabout as the personal car for young people, as well as for wide business usage. Side curtains, opening with doors, starter and four balloon tires are standard equipment.

$360 F.O.B. DETROIT

Fig. 5. Model T Ford advert circa 1925

3 Contradictions and the Wider Impact on Society

The positive impacts of consumerism are offset by some negative effects – which in this paper we call contradictions (although strictly they are more ‘ironies’ than ‘contradictions’). Our concern here is primarily those contradictions related to safety. Consumerism has led to improvements in safety, considering the products or services in themselves, but with negative safety effects if we look more widely. We focus on two environmental impacts – global warming and air quality in cities.

3.1 Global Warming

There is growing evidence that human activity is a major contributor to global warming. There are many factors, including the generation of electricity from fossil fuels, but our focus is transport. The International Panel on Climate Change (IPCC) recently stated ‘carbon dioxide (CO_2) is the main driver of climate change, even as other greenhouse gases and air pollutants also affect the climate’; tourism contributes about 8% of emitted CO_2 and air travel is about half of that (IPCC 2021). Whilst the data is not so conclusive, there are negative effects of emission at altitude, e.g. NO_2 from aircraft. Road transport contributes significantly to CO_2 emissions with around 10% from freight and 15% from passenger transport (Ritchie 2021).[2]

The impact of global warming on human health and safety is evident from the growing frequency of extreme events, including recent destructive hurricanes which have affected the USA. Furthermore, predictions of sea level rise suggest that at least 200 million people will be living below sea level by the end of the century and as many as 630 million are projected to live below annual flood levels on a high emissions scenario (Kulp and Strauss 2019). Alarming as they are, these predictions do not include the impact of melting of the ice shelf.

Thus, although aircraft and cars are remarkably safe in themselves, and becoming more so due to the forces of consumerism, amongst other things, they both contribute to global warming, which has the potential to be an existential threat to humans and many other species.

3.2 Air Quality in Cities

Emissions from road vehicles have an impact on air quality, especially in cities, which, in turn, has an impact on human health. The cause-and-effect relationships between pollution and health are complex to establish. The Committee on the Medical Effects of Air Pollutants points out the dissenting views in the committee on these causal questions (COMEAP 2018).

The COMEAP report presents a detailed analysis of the effect of NO_2 and particulate matter (PM) on premature mortality, seeking to adjust for the correlations between NO_2 and PM, as both can be produced in vehicle emissions. The report contains the estimate that premature deaths in the UK in 2013 were in the range 22,000 to 36,000. The report aims to be balanced and includes questions from dissenters about the (evidence for) causal relationships between NO_2 and premature mortality, however the authors all agree that a reduction of NO_2 and PM will be beneficial to health. The report also considers counterfactuals – in this case, baseline levels of NO_2 and PM against which measured pollution is compared. There are some detailed studies of the effects of PM that identify significant levels of premature mortality in international cohort studies (Burnett et al 2018). Further, there has recently been a conclusion by a coroner that emissions contributed to the death of a schoolgirl in London (The Sunday Times 2020).

As with climate change there is a contrast, or contradiction, between the safety of road vehicles, in themselves, and the wider and longer-term impact of their use. The 1,752 fatalities on UK roads in 2019 due to vehicle accidents is less than

[2] N.B. The figures from (Ritchie 2021) are scaled to be consistent with the IPCC report, but there is not an exact match so the figures should be taken to be indicative not absolute

a tenth of the estimated 22,000 to 36,000 premature fatalities due to poor air quality. As noted above, if electric vehicles had remained a major form of transport in cities as they were over a century ago, then this impact would have been much reduced.

4 Re-framing System Safety Engineering

To resolve contradictions such as the indirect negative safety impact from environmental damage of heavy road and air traffic, we need to broaden safety engineering beyond the normal (narrowly defined) consequences of hazards. Further, we need to include benefits as well as harms and to consider trade-offs between potentially incommensurate benefits and harms. We build up to the re-framing of safety engineering and assurance progressively in the rest of this section.

4.1 The Trade Space

The first step is to define what we call the “trade space” – the range of anticipated impacts that will need to be included in any risk assessment and when evaluating trade-offs. There is an obvious trade-off between the two aspects of consumerism – having fewer cars, aircraft, etc. reduces availability of goods and services, but promotes the interests of consumers in terms of reducing the risks to health and safety arising from adverse environmental impacts. Staying with cars as an example, reducing their availability might mean that an individual travels more by bicycle – at one level, this is good for their health, but at another level they become a VRU and are at about 25 times greater risk per mile travelled than car occupants (Department for Transport 2020).

There will be many other benefits and harms associated with a given product or service – for the owner/user, for the designer or manufacturer, for people directly affected by the system, e.g. those in a city where a car is used, and for society in general. For example, using electric vehicles in a city is beneficial in terms of pollution and hence air quality – but this may just displace pollution rather than reduce it, depending on how the electricity is generated. If it is produced using fossil fuels, then there may be a similar amount of pollution, just in a different place. There are, of course, ethical questions about the acceptability of actions which shift risks (and benefits) between different groups. We identify this as an open question (or future work) in Section 5. It is also worth noting that there is a substantial environmental impact from making cars; whilst there are varying analyses, some suggest that manufacturing and driving a car have similar carbon footprints (Berners-Lee and Clark 2010), and manufacturing electric vehicles has a greater impact than making a comparable conventional vehicle. However, those employed at the factories gain benefits as well as being exposed to the localised risks so, again, there are trade-offs.

In practice, many of these factors are incommensurable. For example, the benefits of being employed at a car factory and the quality of life arising from paid work (psychological, societal), potential harms from the manufacturing processes (physical), and long-term impact from environmental damage cannot obviously be measured or evaluated on a single scale. In addition, given the nature of the supply chain for cars, the risks are quite widely distributed – in mines and quarries, in electronics factories, and so on – making it very hard to calculate risks and to undertake systematic risk-benefit trade-offs. Thus, the first issue is how to re-frame safety engineering to provide a "trade-space" which can be thought of as providing a framework of the different factors – benefits and harms – that need to be considered. We approach this from the viewpoint of well-being.

4.2 Focusing Safety Engineering on 'Well-Being'

The concept of 'dependability' has been long-used to embrace failure-related system properties – safety, availability, reliability, etc. (Avizienis 2004). The models underlying this concept assist with reasoning about the relationship of these key system properties, but dependability does not cover the wider impact on society considered in section 3, and the enlarged trade space described above. The concept of dependability is too narrow, but what is a suitable alternative? We believe that a human-focused approach is essential. Safety is about protecting people from harm. But in the face of contradictions and ironies of long-term negative effects on safety and human well-being from products and services that are "safe in themselves", it seems clear that safety engineering needs to evolve, and to be re-framed to consider not just individual but also societal and environmental impacts.

In philosophy, 'well-being' is what makes life good for the individual living that life – or how well a person's life is going for them (Crisp 2021). It is common for philosophers to draw a distinction between subjective and objective conceptions of well-being or welfare: broadly, whether the concept should be understood in terms of people's own preferences and accounts of what makes life good for them, or in terms of what objectively makes their life go well for them irrespective of their personal predilections. Hybrid theories combine objective and subjective elements of well-being (Parfit 1984). We seek to abstract the following discussion from a commitment to a specific theory of well-being. But any model of well-being that is applied in a re-framing of safety engineering will need to be to some degree objective and codified. The aim is for rational, repeatable safety engineering processes.

One policy approach to well-being with philosophical roots is the capabilities approach – that people need certain capabilities to function well; this derives from Aristotle's notion of *eudaimonia*, or flourishing, as the goal for humans (Nussbaum and Sen 1993). Developing this perspective, many policy-focused analyses decompose the notion of well-being. Some do so on the basis of needs, with, for example, health and personal autonomy taken to be primary, supported by secondary attributes such as nutritional food and clean water, adequate housing, a safe work environment, physical security, and so on (Doyal and Gough 2011).

Another interesting perspective is from Buddhist economics (Schumaker 1966). This considers wider impacts following the use of a product or service, investigating how trends affect individuals, society, and the environment, and links particularly well to the concerns introduced in section 3.

But how can we use the concept of well-being as a basis for enlarged analysis of system safety? We propose a two-layer model. The top level would consider, for a given or proposed system, the potential benefits and harms to individual well-being, society, and the environment. The identified concerns at this level would scope more detailed, lower-level analysis, for example informed by secondary attributes (Doyal and Gough 2011), for identifying benefits and harms (forms of hazard) in sufficient detail so trade-offs and tensions can be considered, and controls defined.

The top level is captured in Table 1. It focuses on impacts at the system level.

Table 1: Categories of Benefit and Harm

Benefits	**Hazards or Harms**
Individual/personal - Physical - Psychological - …	Individual/personal - Physical - Psychological - …
Societal	Societal
Environmental	Environmental

Physical impacts on individuals include improvements to physical safety as well as loss of life or bodily injuries. Psychological impacts include benefits to mental health, and hazards such as addiction and trauma. Societal impacts include benefits and harms to infrastructure and societal functioning (Hassel and Cedegren 2021). Societal impacts also arise from changes to risk distributions. Environmental impacts include issues such as air and water quality and it might be argued that loss of biodiversity has an impact on psychological well-being. In addition, how widely deployed the system is has societal and environmental implications (see 4.3 below). The scope of impact, which we consider under the broad term 'well-being', enlarges safety engineering, both as a discipline and a concern.

4.3 The Numbers Game

After defining the “trade space’ and refocusing safety engineering on an enlarged conception of well-being, it remains to consider the impacts and hazards of widely deployed systems beyond the immediate and discrete impact on individuals. This ties into societal and environmental concerns. Safety engineering normally focuses on a single product or system. By way of illustration, we consider aviation. Safety targets are typically related to hazard classes, e.g. an occurrence rate of $< 10^{-9}$ per flight hour for catastrophic hazards. Such targets apply whether there are only a few aircraft of the type, e.g. Concorde, or a very widely deployed system, e.g. Boeing 737s. When we consider environmental hazards, aircraft-for-aircraft, Concorde would have had a greater environmental impact than an individual Boeing 737 or an Airbus 320. But since there are around 5,000 each of the 737 and 320 in service, their cumulative impact is much greater. In the early days of aviation, environmental impacts were a relatively minor concern. There were very small numbers of aircraft and accident rates were high, so a focus on the direct hazards to occupants made sense. As the analysis in Section 3.2 shows, this is no longer the case, and the sheer volume of air traffic contributes to global warming and thus poses an (indirect) safety risk. So, the next step is that some of the harms (and benefits) need to be considered for whole fleets, not just for individual systems.

Fig. 6. Concorde Leaving New York

4.4 The Interconnectedness of Benefits and Harms

It is also necessary to understand the dependencies and relationships between the different benefits, hazards, and concerns. Over time, regulations have been introduced to address environmental impacts of aircraft (including noise as well as emissions) but again these tend to be at the level of individual aircraft and are disjoint from other safety requirements. Initiatives such as "Net Zero"[3] take a more holistic approach to managing emissions, but not integrating different perspectives such as flight safety with environmental impact. This lack of integration makes it hard to balance different safety concerns. Therefore, it is necessary to consider the dependencies between the different safety concerns to manage them effectively, including indirect hazards to safety from other impacts, such as environmental damage. This consideration will include making trade-offs between both direct and indirect risk related to the same kind of hazard (e.g. to the physical safety of individuals) as well trade-offs between different kinds of hazard (e.g. safety and privacy).

4.5 Safety Engineering Re-framed: Motivational Examples

Safety processes normally start with Hazard and Risk Analysis (HARA). To take this broader view of safety (re-framing it) we suggest a precursor analysis using the notion of benefits and harms to individuals, society, and the environment to scope the issues to be addressed in HARA. We illustrate this by means of two examples.

The CHIRON project is developing a social care robot (see Figure 7), intended to help the elderly and infirm to stand, and thus to continue living independently. The safety of this system has been investigated with funding from the Assuring Autonomy International Programme (AAIP)[4] and this work identified some concerns beyond classical safety issues. These are identified (and amplified) below:

1) Individual
    a) Benefit – enhanced/continued independence (psychological).
    b) Harm – injury from fall (physical); loss of strength/capability over time due to system providing excessive assistance (physical); reduced mental health due to isolation (psychological).
2) Society
    a) Benefit – reduced demand on social care system.

[3] See: https://www.gov.uk/government/publications/net-zero-strategy

[4] See: https://www.york.ac.uk/assuring-autonomy/projects/assistive-robots-healthcare/

   b) Harm – growth in numbers of isolated elderly/infirm individuals later requiring mental health or other support.
3) Environment
   a) Benefit – minimal.
   b) Harm – minimal.

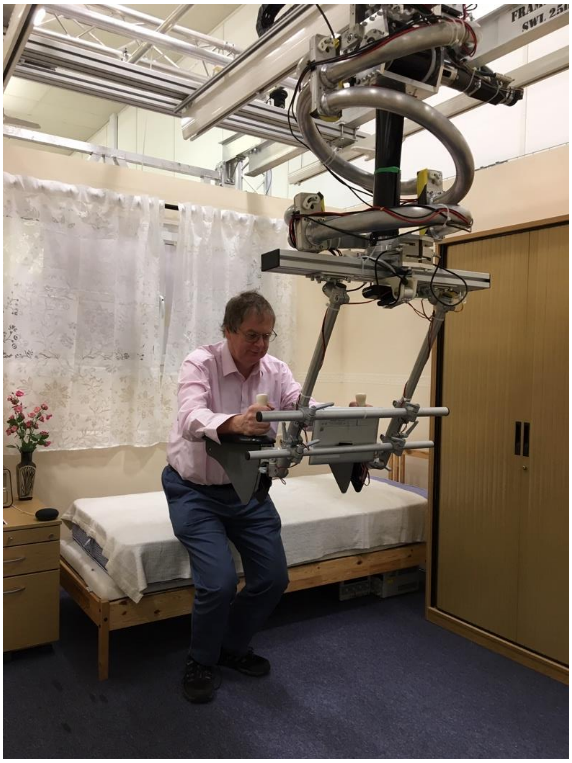


Fig. 7. The CHIRON robot

Traditional safety engineering would address injury from falls, but the other issues require multidisciplinary input, e.g. from physiology and sociology. Broader models, e.g. of the social care system, are also needed for a complete analysis. There will potentially be environmental impact from developing the system, but this is assumed to be minimal as the number of systems is likely to be limited (see the next example for a discussion of supply chain impacts).

The UK has committed to phasing out (new) petrol and diesel cars by 2030. To give more focus, and noting the fact that safety analysis normally addresses particular products or services, we consider delivery vehicles (e.g. developed by

Arrival)[5] but without autonomy, i.e. we assume that the vehicles have a human driver. Here the primary individuals affected are the delivery drivers and those working in factories producing the vehicles:

1) Individual
   a) Benefit – improved air quality for drivers (physical); reduced exposure to hazards from factory automation[6] (physical).
   b) Harm – injury from battery fires (Chen et al 2021) (physical); injury from handling toxic materials in factories (physical).
2) Society
   a) Benefit – ability to deliver products to cities without adversely impacting air quality; enhanced employment and economic benefits for the UK (VividEconomics 2018).
   b) Harm – no obvious issues.
3) Environment
   a) Benefit – aluminium and thermoplastic construction means that the van is light (so there is less impact on the road) and it is more sustainable meaning less impact from re-manufacturing; potential for reduction in pollution long-term through recycling, etc.
   b) Harm – impact in the supply chain from mining for the materials needed for batteries, and their transportation; impact of disposal of batteries at the end of life (Kang et al 2013).

In environmental terms there are also potential impacts of shifts in where the energy is generated, see the discussion in section 3.2. Due to the nature of the system, a long-term view needs to be taken, e.g. of environmental impact and harm.

4.6 Safety Engineering Re-framed: Towards a Process

The AAIP is developing revised safety processes for autonomous systems and robots, including a phase referred to as Societal Acceptability of Autonomous Systems (SOCA). We see the sorts of issues raised here as informing how a broader range of ethical and societal impacts can be incorporated into the assurance framework, but SOCA will also need to address the transfer of decision-making responsibility from humans to machines – and the distinct challenge that autonomy brings.

[5] See: https://arrival.com/?topic=products

[6] See: https://www.wired.co.uk/article/arrival-electric-vehicles-microfactory

As indicated above, subsequent analyses need to be informed by this scoping. We envisage enhancements of classical analysis techniques, e.g. HAZOP, for such stages but discussion of such techniques is outside the scope of this paper.

It is common, in many industries, to produce assurance cases to support decisions to approve a system for use. If assurance cases continue to be used then they need to be expanded to address impacts on society and the environment, as well as to the individuals directly (and indirectly) affected by the system, perhaps ultimately taking a 'Global Safety' perspective. There will also be a need to reason about benefits vs harms (risks). In current practice, arguments of risk versus benefit, including considering costs of options, are carried out where there is difficulty in reaching risk targets and the developers wish to show that risks are reduced as low as reasonably practicable (ALARP). Given the re-framing of safety we propose here then the arguments of benefit versus harm needs to be considered in all cases, not just *in extremis*, to seek to identify an "optimal" system concept and implementation. Here, the "costs" will need to include harms to society and environment, and not just focus on engineering economics. In principle it is (legally) necessary to produce ALARP arguments comparing all possible designs. There is a major problem with ALARP arguments which becomes worse in the re-framed safety process – and here we see a role for counterfactual thinking.

4.7 Safety Engineering Re-framed: Counterfactuals through Life

In Section 2, we indicated how counterfactual thinking can contribute to reasoning about longer-term and indirect impacts to human well-being from high-integrity systems. Counterfactual explanations have been adopted by the AI and machine learning (ML) communities as a way of explaining the behaviour of otherwise opaque algorithms (Wachter et al 2018). They are a form of example-based reasoning (Miller 2019). They are likely to have a role in safety assurance of systems and products as ML becomes more widely used and interpretability of decisions made by complex systems becomes necessary (Jia et al 2021a). But we also see a broader role for counterfactuals which we consider as part of our proposed re-framing of system safety engineering.

It is often stated that safety engineering is a through-life concern but, in practice, much of the emphasis is on pre-deployment analysis of a particular system – with the assumption that systems will remain safe (enough) through life if well-maintained and operated appropriately. It is suggested here that the through-life nature of safety engineering needs to be widened and reinforced to consider very early life-cycle issues (concept design), and through-life system monitoring, e.g.

the lifetime environmental impact from a vehicle, including disposal, as well as accident and incident analysis.

First, is to incorporate "counterfactual thinking" at the earliest stages of system design, considering major design options. For example, if road infrastructure is modified to include inductive charging, then electric vehicle battery sizes could be reduced. There is an environmental impact (harm) in reworking existing roads and the implications for embodied carbon[7] to set against the benefits of reduced battery requirements, and the consequent change to the supply chain (including factories). By considering "what is not but could have been" for the major options it will give a basis for reasoning about the "best" alternative.

Practical trade-offs tend to compare possible changes to a baseline design – a similar approach could be adopted here, looking at the "delta" in benefits and harms in comparing the proposed design with alternatives (this is consistent with the way counterfactuals are generated in ML, seeking to minimise the change in inputs to produce the desired outputs). These trade-offs would need take a through life perspective, and it may be that approaches from economics or healthcare, such as Quality Adjusted Life Years (QALYs)[8], would be useful. Given the breadth of considerations in this re-framing of safety engineering, there will be many stakeholders – and the majority of these will be "lay", in the sense of not being specialist in the technology. Again, we would see the counterfactuals as having a role – "we don't recommend this option because …". However, the environment can't "speak for itself" and there will be a need to seek out appropriate stakeholders covering all the relevant concerns.

Second, it is important to monitor systems in operation to see whether or not they behave as predicted – including in safety terms. A number of projects are exploring the notion of learning from operations, including through use of digital twins, for example at the Alan Turing Institute[9]. It is possible to apply ML to operational data to identify cases where system behaviour deviates from what was predicted in a way that has implications for safety (Jia et al 2021b). Further, with very complex systems it is hard to predict all the possible behaviours of the system in advance and its wider impacts in terms of society and the environment (McDermid et al 2021). Here the need is to monitor for continuous changes/long-term trends, not events. We suspect that counterfactual approaches could help here – for example identifying the minimum set of changes necessary in the system or its operation to achieve the intended balance between benefits and harms.

Third, effective learning from accidents and incidents includes a form of counterfactual reasoning. Section 2.1 illustrated the long-term safety improvements that have been achieved in air transport. One of the reasons for this success is the

[7] See: https://www.raeng.org.uk/RAE/media/General/Policy/Net%20Zero/NEPC-Policy-Report_Decarbonising-Construction_building-a-new-net-zero-industry_20210923.pdf

[8] See: https://www.nice.org.uk/glossary?letter=q

[9] See: https://www.turing.ac.uk/research/research-projects/theoretical-foundations-engineering-digital-twins

thoroughness of accident investigation in seeking the underlying root causes of accidents, not just the proximate cause. The above-mentioned emphasis on CRM (crew/ cockpit resource management) is a case in point. Although not usually expressed this way, learning from experience in safety is concerned with finding actionable counterfactuals – if x had not occurred, y would not have occurred; therefore, we need to remove the possibility of x or reduce its probability to avoid occurrence of an accident with a particular signature.

In addition, many modern systems are data rich. It may therefore be possible to see the chain of events that led up to an accident in the data and to use explainable AI (XAI) methods to generate counterfactuals during analysis of accidents and incidents – identifying the minimum changes that could have prevented the accident. Care is needed, however, in that ML identifies correlations in data, not causation, so all learnt models and suggestions need to be subject to human scrutiny. For example, a counterfactual that says an aircraft would have avoided a runway overrun if it hadn't landed is true but unhelpful. One interesting area of work is on contrastive explanations (Lipton 1990), including identifying "pertinent negatives": factors whose *absence* is necessary to draw a particular conclusion (Dhurandar et al 2018). A contrastive explanation that compares accident-free behaviours with accident scenarios might identify missing controls – pertinent negatives – that, if present, could have prevented the accident from arising. Note that this will not work in all cases. If the controls needed are new – not already an aspect of system design or operation – then this approach will not find them. Counterfactual thinking is still valuable, but automated ways of generating counterfactuals will not be a panacea.

Finally, whilst we have indicated a role for ML in supporting system monitoring and accident/incident analysis, we note that the data centres that support on-line services, including ML applications, account for about 1% of global electricity supply[10]. Whilst this is not all due to ML, this environmental impact needs to be considered and suggests, again, the need to think holistically. The reframed safety engineering framework, focused on a broad notion of 'well-being' – and with a central role for counterfactual explanations, analysis, and reasoning – ought to be a powerful tool in shaping policy and choosing amongst policy options, including allowing for their longer-term impact.

5 Discussion and Conclusions

In this paper we have striven to be bold yet realistic! We have highlighted some challenges for a modern conception of system safety, starting from the contradictions of consumerism. Increased availability of safe products and services, for

[10] See: https://www.iea.org/reports/data-centres-and-data-transmission-networks

example cars and air travel, leads to environmental damage — and ultimately deleterious effects on human safety, health, and well-being. In light of such concerns, we have suggested steps for a progressive re-framing of safety engineering. This adopts the wider goal of 'well-being' as part of a more holistic, human-centred approach. The shift in focus covers societal and environmental impacts, as well as the impact both on individuals' psychological and physical well-being. Although the scope is wide, we believe our discussion shows the merit in a simple overarching structure to frame the concerns.

While there is other work on re-framing safety engineering, we don't believe any have the necessary scope to deal with the broad issues identified here. Safety II (Hollnagel 2018) focuses on complex socio-technical systems and on understanding "what goes right" as well as considering failures (which he characterises as Safety I). Undoubtedly this is a useful mindset, but it doesn't address the wider societal and environmental issues identified here.

Our ideas can be seen as endorsing the inclusion of a wider ethical perspective in the evolution of safety engineering. There are several initiatives around ethics of autonomy, for example the IEEE's work on Ethically Aligned Design and their P7000 series of standards.[11] The P7000 initiative is important and visionary, and includes documents addressing well-being (IEEE 2020), but the work is restricted to autonomous and "intelligent" systems. As should be clear from the above examples, the scope of concern here is much broader. The approach we have outlined should give a basis for realising ethically aligned design on a broader front than just autonomous systems – although that remains to be demonstrated.

The re-framing of safety engineering that has been suggested in this paper is broad. If it can be realised at all, then it can only be achieved over time. Some of the more speculative ideas need to be assessed for feasibility and we see this paper as the "start of a journey" not a well-specified destination. There are also open questions for future work and systematic reflection which should guide that journey:

1. Can safety engineers address all these concerns (individual, societal, environmental impact) alone? If so, how do they achieve the necessary knowledge? Alternatively, should safety engineers take on a role of integrating thinking from a range of disciplines, perhaps with an emphasis on articulating the trade-offs between incommensurable concerns?
2. How should we approach defining the objective (at least calculable) measures of well-being and acceptable risk? Again, there is a need to draw on the theoretical resources of other disciplines, perhaps econometrics or healthcare (e.g. adopting or adapting the QALY) but can general risk classes be defined, or do risks need to be evaluated on a case-by-case – system-by-system – basis?

[11] For the list of P7000 standards can be found see: https://ethicsinaction.ieee.org/p7000/

3. How can we reason about distributions in risks (harms) and benefits from (new) systems, and how do we engage relevant stakeholders in decision-making? This is likely to need methods of stakeholder engagement from social science. Counterfactual explanations might also be useful in communicating about alternative possibilities to a diverse, and lay, audience.

In our view, the need for re-framing system safety is pressing. The ideas presented here are intended to stimulate debate and help to influence the future evolution of safety engineering. One thing is clear; this must be a multidisciplinary undertaking. Ethicists, psychologists, environmentalists, human factors specialists, and experts in supply chains all need to be involved, almost regardless of the system considered. It is also likely that data scientists and experts in AI/ML will make a substantial contribution to developing practicable new analysis methods, especially when considering operational data. In specific domains, e.g. aviation, other specialists, e.g. atmospheric chemists, will need to be involved. Our hope is that we have provided enough of a starting point to enable these disparate groups to start to work together within an enlarged, human-centred framework for safety engineering.

Acknowledgements This work was supported by the Assuring Autonomy International Programme, funded by the Lloyd's Register Foundation and the University of York.

References

Automobile Association (AA) (2021) From windscreen wipers to crash tests and pedestrian protection https://www.theaa.com/breakdown-cover/advice/evolution-of-car-safety-features. Accessed 11 October 2021

Avizienis A, Laprie J-C, Randell B and Landwehr C (2004) Basic concepts and taxonomy of dependable and secure computing, in IEEE Transactions on Dependable and Secure Computing 1(1):11-33,

Berners-Lee M and Clark D (2010) What's the carbon footprint of ... a new car? https://www.theguardian.com/environment/green-living-blog/2010/sep/23/carbon-footprint-new-car. Accessed 7 September 2021

Boeing (2020) Statistical summary of commercial jet airplane accidents: worldwide operations, 1959-2020. https://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf. Accessed 11 October 2021

Burnett R et al (2018) Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences 115(38): 9592-9597

Chen Y et al (2021) A review of lithium-ion battery safety concerns: The issues, strategies, and testing standards. Journal of Energy Chemistry 59: 83-9

COMEP (2018) Associations of long-term average concentrations of nitrogen dioxide with mortality, Public Health England 2018238

Crisp R (2021) Well-Being, The Stanford Encyclopaedia of Philosophy (Fall 2021 Edition), Zalta E (ed.). https://plato.stanford.edu/archives/fall2021/entries/well-being. Accessed 11 October 2021

Department of Energy, U.S. (2014). The history of the electric car. https://www.energy.gov/articles/history-electric-car. Accessed 11 October 2021

Department for Transport, G.B. (2020) Reported road casualties in Great Britain: 2019 annual report https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/922717/reported-road-casualties-annual-report-2019.pdf. Accessed 11 October 2021

Doyal L and Gough I (2011) A theory of human needs. MacMillan

Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K. and Das, P. (2018) Explanations based on the missing: towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623.

Hassel H and Cedergren A (2021). A framework for evaluating societal safety interventions, Safety Science, 142:105393

Hollnagel E. (2018). Safety-I and Safety-II: the past and future of safety management. CRC press.

IEEE (2019) Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html. Accessed 11 October 2021

IEEE (2020) IEEE P7010-2020, IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. https://standards.ieee.org/standard/7010-2020.html. Accessed 11 October 2011

IPCC (2021) Climate change – widespread, rapid and intensifying, https://www.ipcc.ch/2021/08/09/ar6-wg1-20210809-pr/. Accessed 6 September 2021

Jia Y, McDermid, J A, Lawton T and Habli I (2021a) The role of explainability in assuring safety of machine learning in healthcare. Submitted to IEEE Transactions on Emerging Topics in Computing (available at: arXiv preprint arXiv:2109.00520).

Jia,Y, Lawton T, McDermid J A, Rojas E and Habli I (2021b) A framework for assurance of medication safety using machine learning. arXiv preprint arXiv:2101.05620.

Kang D, Chen M, Ogunseitan O (2013) Potential environmental and human health impacts of rechargeable lithium batteries in electronic waste. Environ Sci Technol 47(10):5495-5503.

Kulp S and Strauss B (2019). New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. Nature communications 10(1): 1-12.

Lipton P (1990) Contrastive explanation. Royal Institute of Philosophy Supplements 27:247-266.

Lyth P J, Dierikx M L J (1994) From privilege to popularity: the growth of leisure air travel since 1945. J Transport History 15(2):97-116

McDermid JA, Burton S, Garnett P, Weaver RA (2021) An initial framework for assessing the safety of complex systems, in Parsons M and Nicholson M (eds). Systems and COVID-19: Proceedings of the 29th Safety-Critical Systems Symposium Virtual Conference,

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. 267: 1–38

Nussbaum M and Sen A (1993) Capability and Well-being, in Nussbaum M and Sen A (eds.) The quality of life. Clarendon Press

Owen R, Bessant J. Heintz, M. eds. (2013) Responsible innovation: managing the responsible emergence of science and innovation in society. John Wiley & Sons.

Parfit D (1984) Reasons and persons. Oxford University Press

Ritchie H (2021) Cars, planes, trains: where do CO2 emissions from transport come from? https://ourworldindata.org/co2-emissions-from-transport. Accessed 6 September 2021

Schumaker E F (1966) Buddhist economics in Asia: a handbook, Wint G (ed., Anthony Blond Ltd. https://web.archive.org/web/20121213145110/http://neweconomicsinstitute.org/buddhist-economics Accessed 8 September 2021

Starr W (2021), Counterfactuals, The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.) https://plato.stanford.edu/archives/sum2021/entries/counterfactuals. Accessed 11 October 2021

Statista (2021) Number of scheduled passengers boarded by the global airline industry from 2004 to 2020 https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/. Accessed 6 September 2021

The Sunday Times (2020) Vehicle emissions in the spotlight again as coroner concludes air pollution contributed to death of schoolgirl https://www.driving.co.uk/news/environment/air-pollution-contributed-death-nine-year-old-coroner-rules/. Accessed 7 September 2021

World Bank (2021) Air transport, registered carrier departures worldwide https://data.worldbank.org/indicator/IS.AIR.DPRT. Accessed 11 October 2021

VividEconomics (2018) Accelerating the EV Transition Part 1: environmental and economic impacts. https://www.wwf.org.uk/sites/default/files/2018-03/Final%20-%20WWF%20-%20accelerating%20the%20EV%20transition%20-%20part%201.pdf. Accessed 11 October 2021

Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard Journal of Law & Technology 31(2).

Von Schomberg R (2013) A vision of responsible research and innovation, in Owen R, Bessant J, and Heintz M,(eds) Responsible innovation: managing the responsible emergence of science and innovation in society. John Wiley & Sons